THE LOGIC OF EVALUATIVE ARGUMENT

I choose the word "argument" thoughtfully, for scientific demonstrations, even mathematical proofs, are fundamentally acts of persuasion. Scientific statements can never be certain; they can only be more or less credible.

Joseph Weizenbaum, *Computer Power and Human Reason,* 1976

Generalizations decay.

Lee J. Cronbach, *Beyond the Two Disciplines of Scientific Psychology,* 1974

The Coming Great California Earthquake

I sit in Los Angeles but wonder why I stay. A sudden one-foot uplift has appeared along a hundred-mile strip of the San Andreas fault. Based on seismic wave readings, a California scientist has predicted a major earthquake for the Los Angeles area within a year (*Science,* May 1976). Based on different readings, a radio evangelist warns of a major quake. Both scientists and seers agree in their prophecies. Neither provides the kind of information I need.

I talk to the natives about these ominous signs. Their response is shaped by the necessity of living in such circumstances; they shrug their shoulders. The President has been informed, but no one seems to know exactly what to do. Washington officials suggest setting up a new array of scientific instruments along the fault, although what will result from more measurement is not clear.

AUTHOR'S NOTE: This chapter is a shorter version of *The Logic of Evaluative Argument,* Monograph #7, Los Angeles: Center for the Study of Evaluation, UCLA, 1977. I wish to thank Lee Cronbach, Bob Ennis, Gene Glass, and the CSE staff for detailed comments.

Meanwhile the weather is perfect, the setting in the Santa Monica Mountains splendid, the lifestyle sybaritic. Calculations of probabilities of long-term seismic events do me no good; I need to know when the earth will move in relation to myself. The vocabulary of action is complex. Everyone agrees that information somehow informs decisions, but the relationship is not direct, not simple. Often the more important the decision, the more obscure the relationship seems to be. Consider the decision to marry. For most people, it is a long, arduous process, one which takes shape over a period of time. No single piece of information serves as a decision-point. Quite the contrary. The decision proceeds slowly, almost imperceptibly, until it arrives. Reason after reason is advanced and tried out. Finally, a multiplicity of arguments serves as a rationale for the decision, which is often made long before all the arguments are advanced.

The most significant decisions are those that have long-range implications but defy easy extrapolation, that are so entangled with everything else that they resist precise formal analysis. To those we are forced to apply our intuitive logic, our common sense. It is in the nature of these complex problems that knowledge about them is limited, that it is less than determinate. In the face of uncertain knowledge, the task of entangled decision-making becomes less one of absolutely convincing ourselves with proofs than one of persuading ourselves with multiple reasons. The criterion becomes not what is necessary but what is plausible.

Equivocality of Evidence: Certainty vs. Credibility[1]

Why, then, do government officials, the public, and even members of the evaluation community call for definitive proof of the success of educational programs? There is a tradition as old as Descartes which says that the *only* knowledge is that which is certain. Descartes's method of analysis was one of total skepticism: to doubt everything that could be doubted. In his search for certain knowledge, he arrived at the *self-evident* as the ultimate mark of reason. For something to qualify as knowledge it had to start from clear and distinct ideas and be extended by deductive proofs. Propositions so derived were thus necessary and compelling to the intellect; they could not be rationally denied.

This method excluded the merely credible from consideration as knowledge. In the Cartesian ideal, the only true reasoning is analytic. Formal deductive logic, the method of proof used in mathematics, is the method par excellence. Knowledge can be reduced to self-evident propositions. In certain knowledge there can be no disagreement. As Descartes wrote, if there is disagreement over a matter between two men, one of them must surely be wrong. There is a true and a false, and logic works by compelling proofs to determine which is which.

Later, those who pursued this line of reasoning confronted the fact that rational men often seemed to reason differently and arrive at contradictory conclusions. Some of Descartes's own propositions looked suspicious. Pascal introduced the explanation that such disagreement, as well as the reluctance to accept necessary conclusions, was a result of irrationality. Man was seen to possess an irrational side which often led him astray in his search for knowledge. The apparent irrationality of those who do not accept conclusions which others perceive as compelling is a common motiff in contemporary evaluation.

From the Cartesian perspective, certain knowledge can be obtained by deductive processes, and it must lead to absolute conviction. Such reasoning may work in geometry, but it does so by excluding most of the sensate world. As Hume pointed out, our beliefs, even in concepts as basic as causality, are not certain when a thorough skepticism is applied to them. Deductive reasoning succeeds in producing certain knowledge primarily by eliminating most of the every-day world.

The sensate world was epistemologically salvaged for our use by John Stuart Mill. Just as logicians had constructed formal deductive logic by reflecting on the nature of mathematical proofs, Mill reflected on the associationist psychology of his time and formulated an inductive logic that purported to introduce certainty into inductively derived knowledge. To do this Mill made several assumptions that still pervade survey research today. According to Hamilton (1976), the axioms include the following:

- There is a uniformity of nature in time and space. This lends to inductive reasoning the same procedureal certainty as to conclusions drawn from syllogistic logic.
- Concepts can be defined by direct reference to empirical categories and laws of nature can be inductively derived from data because of the above.

- Large samples can suppress idiosyncracies and reveal "general causes."
- The social and natural sciences have the same aim of discovering general laws (which provide a basis for explanation and predictions).
- The social and natural sciences are methodologically identical.
- The social sciences are merely more complex.

Thus, Mill contended that certain knowledge was derivable from inductive reasoning as well as from the deductive. One could define categories and relate them to each other by now familiar techniques. In fact, Mill concluded that the inductive method was the *only* way of discovering new ideas since deductive logic could only reveal what was already there. (Mill was so certain of his method that he contended that ethical principles could also be derived by inductive reasoning and hence had a scientific base.)[2]

Mill's first assumption is the important one. In Mill's own words, "The Universe, so far as known to us, is so constituted, that whatever is true in one case, is true in all cases of a certain description; the only difficulty is, to find what description" (Mill, 1893). How familiar that idea is to anyone who has engaged in survey research, and how fallible the inductive logic on which it is based!

The procedure of reasoning from "some" to "all" is clearly a logical fallacy. Each confirming instance is supposed to make a hypothesis more likely. Yet if the hypothesis is "All men are less than 100 feet tall" and one finds a man 99 feet, this is a confirming instance that weakens the hypothesis considerably rather than strengthens it (Gardner, 1976). Does every day that goes by in Los Angeles without the predicted great quake make it more or less likely? It is also quite possible in statistical studies to confirm a hypothesis by two independent studies and yet disconfirm the hypothesis by using the total results of the two studies taken together (see Simpson's paradox in Martin Gardner, 1976).

Nonetheless, in spite of serious flaws of logic, "science" based on inductive logic seems to work with some degree of success. Certainty of knowing, however, is lacking. Even the best established scientific facts must be held as tentative. As one scientist put it:

> The man in the street surely believes such scientific facts to be as well-established, as well-proven as his own existence. His certitude

> is an illusion. Nor is the scientist himself immune to the same illusion. In his praxis, he must, after all, suspend disbelief in order to do or think anything at all. He is rather like a theater-goer, who, in order to participate in and understand what is happening on the stage, must for a time pretend to himself that he is witnessing real events. The scientist must believe his working hypothesis, together with its vast underlying structure of theories and assumptions, even if only for the sake of the argument. Often the "argument" extends over his entire lifetime. Gradually he becomes what he at first merely pretended to be: a true believer. I choose the word "argument" thoughtfully, for scientific demonstrations, even mathematical proofs, are fundamentally acts of persuasion.
>
> Scientific statements can never be certain; they can be only more or less credible. And credibility is a term in individual psychology, i.e., a term that has meaning only with respect to an individual observer. To say that some proposition is credible is, after all, to say that it is believed by an agent who is free not to believe it, that is, by an observer who, after exercising judgment and (possibly) intuition, chooses to accept the proposition as worthy of his believing it [Weizenbaum, 1976].

Evaluation as Persuasion

If demonstrations in the physical sciences are fundamentally acts of persuasion, inquiries in education are more so. Mill's assumption that the social and natural sciences are methodologically identical seems much more dubious today. Cronbach (1974), for one, doubts the advisability of imposing physical science ideals in social science. In the physical science paradigm, events are explained and predicted by "a network of propositions connecting abstract constructs."

After reviewing twenty years of aptitude treatment interaction studies, which were based on such a model, Cronbach concluded that social phenomena are too open to interactions with other variables to support stable generalizations. The positivistic strategy of fixing conditions in which to reach generalizations assumes steady processes that can be separated into independent systems for study, a fragile assumption in social systems.

Cronbach has suggested interpreting data in context rather than trying to arrive at generalizations. An observer in a particular setting can describe and interpret effects within local conditions. Whereas experimental control and systematic correlation ask formal questions in advance, local observation is more open to the unanticipated. Short-term empiricism is sensitive to the context. In being context sensitive, the researcher may give up some predictive power. He gives up constructing generalizations and theory-building and instead develops "concepts that will help people use their heads." So Cronbach contends.

Evaluations themselves, I would contend, can be no more than acts of persuasion. Although sometimes evaluators promise Cartesian proof and use J. S. Mill's methods of induction, evaluations inevitably lack the certainty of proof and conclusiveness that the public expects. The definitive evaluation is rare, if it exists at all. Even a scientific methodologist as sophisticated as James Coleman is faced with continued and trenchant criticism of his work. Subjected to serious scrutiny, evaluations always appear equivocal.

Expecting evaluation to provide compelling and necessary conclusions hopes for more than evaluation can deliver. Especially in a pluralistic society, evaluation cannot produce necessary propositions. But if it cannot produce the necessary, it can provide the credible, the plausible, and the probable. Its results are less than certain but still may be useful.

Proving something implies satisfying beyond doubt the understanding of a universal audience with regard to the truth. To produce proof that a universal audience comprising all rational men would accept requires overcoming local or historical particularities. Certainty requires isolating data from its total context as, for example, in the terms of a syllogism. Logical certainty is achievable only within a closed, totally defined system like a game.

If evaluation is limited to certain knowledge provided by strict deductive and inductive reasoning, it must abandon a great amount of reasoning power that people ordinarily use in the conduct of their lives. Such a limitation results from confusing rationality with logic. They are not identical.

If absolutely convincing all rational men is too heavy a burden for evaluation, persuading particular men is not. In place of the compelling propositions derived from rigorous logic, one may substitute the non-compelling arguments of persuasion. In place of the necessity of self-evidence, one may substitute variable adherence to theses as presented to particular audiences. The thesis may be more or less credible. The audience is free to believe or not believe after inspecting the arguments and exercising its own judgment.

Evaluation aims at persuading a particular audience of the worth of something or that something is the case by an appeal to the audience's reason and understanding.[3] For this purpose, uncertain knowledge is useful although the ideas themselves are always arguable. The appropriate methods are those of argumentation, which is the realm of the "credible, the plausible and the probable" rather than the necessary (Perelman and Olbrechts-Tyteca, 1969).

Argumentation is contrasted to demonstration. Demonstrations rest on formal logic which avoids ambiguity by the internal consistency of its symbol system. In deductive logic the origin of the axioms is extraneous. When one moves from deduction to induction, all manner of issues become arguable, such as the validity of measurement. But the search is still for "certain" knowledge.

In evaluation, the social and psychological contexts become particularly relevant and the knowledge less certain. Under those conditions argumentation aimed at gaining the adherence and at increasing the understanding of particular audiences is more appropriate. Persuasion claims validity only for particular audiences and the intensity with which particular audiences accept the evaluative findings is a measure of this effectiveness. The evaluator does not aim at convincing a universal audience of all rational men with the necessity of his conclusions.

Persuasion is directly related to action. Even though evaluation information is less certain than scientific information addressed to a universal audience, persuasion is effective in promoting action because it focuses on a particular audience and musters information with which this audience is concerned. Personalized knowledge that induces people to stop smoking may be different from scientific generalizations linking smoking to heart disease or cancer. Finding out about the heart attack of a close relative is more likely to induce one to exercise than are charts and tables. Evaluative argument is at once less certain, more particularized, more personalized, and more conducive to action than is research information.

In summary, evaluation persuades rather than convinces, argues rather than demonstrates, is credible rather than certain, is variably accepted rather than compelling. This does not mean that it is mere oratory or entirely arbitrary. The fact that it is not limited to deductive and inductive reasoning does not mean that it is irrational. Rationality is not equivalent to logic. Evaluation employs other modes of reasoning. Once the burden of certainty is lifted, the possibilities for informed action are increased rather than decreased.

The Evaluation Audiences

If persuasion becomes the aim of evaluation, the audiences to whom the evaluation is addressed are important. For years evaluators have been counseled to think of their audiences and the kind of information the audiences will need. What is relevant for one group may not be relevant for another. Argumentation presupposes that a "community of minds" exists, that there is intellectual contact, and that there is agreement on at least a few issues on which deliberation is to begin.

There must be a common language and a desire on the part of the evaluator to persuade the audiences and to take their concerns seriously. Often these conditions are not met. The audiences are misconceived or not taken seriously. It is not uncommon for the evaluator to muster information appropriate to an audience of psychologists but which has little meaning for a teacher or a government official.

The agreement of a universal audience (all men at all times) is likely to be secured by formal logical reasoning based on self-evident concepts. Thus the tighter the experimental design, the more convinced a far-removed universal audience wil be of the cause and effect relationship, regardless of the context. A particular audience closer to the scene may assume cause and effect without such proof. Of course, the universal audience is not "aggregatable" at any given time, but various elite groups in fact serve as a surrogate for it. Perhaps philosophers more than most represent this type of audience. The arguments that move philosophers are not always the same as those that move teachers.

The more an argument is directed toward a universal audience, the less "arguable" it is. There is little to argue about in pure deductive logic. Evaluation techniques are often presented as being nonargumentative, as, for example, being based on valid and reliable instruments, as employing sound statistical procedures, and so on. In fact, all statements made on the basis of an evaluation are subject to challenge and are arguable—if properly challenged. The more technical and quantitative the evaluation, the less a naive audience will be able to challenge it, and the evaluation will appear to be more certain than it is.

In evaluations using statistical metaphors, one can argue that treatment effects differ because there is a probability that two mean test scores belong to different populations and, hence, that the experimental program is better than the control. The extensive use of numbers in the statistical procedures and the test scores gives a semblance of certainty and unequivocality to evidence.

Actually, many assumptions lie concealed behind the numbers (as indeed behind every evaluation). One can almost always challenge the validity of the tests, the appropriateness of the statistical procedures, and the control of the experimental design. The challenge does not invalidate the evaluation. But once the premises are challenged, the nature of the evaluation as argumentation becomes apparent. The evaluator may defend his study either successfully or unsuccessfully. In any case, he must resort to nondeductive and more equivocal reasoning if he is to defend it. Although the evaluation has the appearance of appealing to the definitive rationality of the universal audience, it ends in direct appeals to particular audiences. I believe it is impossible to construct an evaluation otherwise.

Even a broad-based evaluation operation like *Consumers Report,* which uses "objective" procedures and sophisticated experimental designs to evaluate consumer products, is an appeal to particular audiences. Its arguments, directed at the upper-middle class, have little meaning for either the lower classes or the upper classes, and its evaluations are little heeded by them.

Thus the situation the evaluator faces is almost always an appeal to particular audiences which he can define with some precision. If he cannot define his audiences, the evaluation is indeterminate. He must address issues and construct arguments that appeal to particular audiences. Furthermore, the audiences are likely to be a composite of several groups, which complicates his task considerably. Effective appeal to particular audiences changes the limits of applicable rationality. One is not confined to the most restrictive modes of reasoning. If evaluation becomes more equivocal, it also becomes more possible.

One ideal of two-party argumentation is embodied in the Socratic dialogue. The dialogue develops as a rigorous chain of reasoning between a questioner and a responder. The one-person audience is persuaded by getting her to agree on certain principles point by point. The audience's particular concerns are ultimately addressed in the interaction. The Socratic dialogue is also powerful to third parties who might read it.

The actual audience most evaluators face seldom consists of one person, however. It is most often several different groups. Some evaluation theorists have suggested modes of evaluation in which the evaluator engages in frequent exchange with the audience throughout the study. Whatever the mode of evaluation, I would contend that evaluation which succeeds in being persuasive must engage the audience in fundamental discourse, although that discourse may occur in different ways.

Discourse conducted in this fashion is more than a mere debate in which different points of view are presented by partisans. The dialogue must be a discussion in which the parties seriously and honestly search for mutual answers. This restriction severely qualifies the use of adversary methods as persuasive devices since one may adjudicate a conflict without persuading anyone of anything.

Kemmis (1976) has advocated oneself as the audience—"evaluation as self-criticism." He sees the primary audience as being the program staff itself. Believing a dialectic between knowledge and action to be the only way to improve practice, he has suggested that evaluation standards be derived from the program participants themselves and that the data consist of the progress as seen by participants. Evaluation thus becomes therapeutic self-criticism. The ultimate goal is increased understanding and insight of the participants themselves, which can then lead to effective action.

Whoever the audience, in argumentation, the audience must share responsibility. Since the information is not compelling, the audience is free to choose its own degree of commitment. It must actively choose how much it wishes to believe. This requires an active testing of the evaluation by the audience itself rather than a passive acceptance or rejection. The audience must make a personal commitment and share responsibility. This rational decision belongs to the audience, not to the evaluator.

Premises of Agreement

The development of an evaluation argument presupposes agreement on the part of the audiences. The premises of the argument are the beginning of this agreement and the point from which larger agreement is built. Just as common sense admits unquestioned truths that are beyond discussion, some of the major premises of an evaluation are tacit rather than explicit.

According to Perelman and Olbrechts-Tyteca (1969), there are two classes of premises: the "real" and the "preferable." The real includes facts, truths, and presumptions and generally claims validity vis-à-vis the universal audience. On the other hand, the preferable is identified with a particular audience and includes values, composite value hierarchies, and value premises of a very general nature called "loci."

Facts and truths are those data and notions which are seen as agreed upon by the universal audience, i.e., held in common by thinking beings, and hence needing no justification. Whether a datum is a fact depends upon one's conception of the universal audience. If the audience changes, so can facts and truths. However to hold the status of a fact or a truth means that for the purposes of argument the datum is noncontroversial and uncontested. If the datum is questioned, it loses its status as a fact and becomes itself an object of argument rather than an object of agreement.

Where there is agreement on the conditions for verification as in modern science, there can be many facts. Many data are not accorded the status of "facts" by modern science. Polanyi (1958) pointed out how science protects its own system of beliefs from inconsistency by denying that various data which conflict with other beliefs are factual. Thus for many years science did not recognize hypnotic effects as occurring at all. These data were not recognized as factual because they conflicted with the current general scientific belief system. This belief system may change from time to time, but regardless of what it excludes, arguments within the belief system must be based on uncontested facts and truths.

Arguments also proceed from presumptions which do not have the full authority and confidence of a fact or truth. Presumptions cannot be proved but are nonetheless widely accepted as being tentatively true. Many presumptions are connected to the concept of the normal. In evaluations employing statistical models and metaphors, the assumption that attributes within a population are normally distributed is almost universally accepted.

The second class of objects of agreement is that of the preferable. Objects of preference claim the adherence of only particular groups rather than that of the universal audience. Values are the most conspicuous examples. Agreement with regard to a value is an admission that there is a specific influence on action or a disposition toward action that the evaluator can make use of. Although relevant for a particular group, a value is not regarded as binding on everyone.

In science, values enter primarily in the selection of objects of interest for investigation since one cannot investigate the entire world (Polyanyi, 1958) and possibly in the acceptance of scientific conclusions by overall human judgment (Weizenbaum, 1976). But during most of the argument, especially in the exact sciences, values are supposed to be excluded. Ennis's (1973) analysis of cause and effect relationships leads one to question this. In evaluation there is no question that values enter at every stage. Values are used to persuade the audiences and to justify choices to others.

Various combinations of arguments can be compressed into a few general groupings called "loci" (Perelman and Olbrechts-Tyteca, 1969). The most common loci are those of quantity and quality. Arguments grouped around the loci of quantity affirm that one thing is better than another for quantitative reasons–greater number, higher degree, more durability, etc. The effectiveness of means will often be justified by quantitative loci. The idea of the normal and the norm are also based on quantity.

Contrasted with quantity is the idea of quality. Something has high value even though it defies number. Associated with quality is a high rating of the unique. One can be in possession of truth while the multitude is in error. For example, Scriven (1972) contended that the notion of objectivity is not necessarily linked to the number of people holding an idea, nor subjectivity to one person's perception, as is often believed.

Besides general agreements on facts and values, there are special agreements particular to certain special audiences and particular to each evaluation. To the extent that the evaluation is addressed to a technical audience, that audience will share certain agreements and conventions. A group of educational researchers is such a technical audience. Evaluations directed toward a lay audience cannot rely on the same agreements.

Perhaps the most important agreements peculiar to a particular evaluation are those derived from the negotiation that often precedes the evaluation–agreements between sponsors, program personnel, and evaluators. In this exceedingly important negotiation, agreement can be reached on criteria, methods and procedures, access, dissemination of results, and so on. Disagreement on these points can destroy the entire credibility of the evaluation.

In summary, at the beginning of an evaluation, the evaluator must build upon agreements with the audiences. These agreements may be implicit as well as explicit. In fact, it would be impossible to specify all these understandings, although it is dangerous to assume agreement on important points where there is none. The evaluator must start from where his audiences are, even though the beginning premises may not be acceptable to other parties nor to the evaluator himself. Otherwise the evaluation will not be credible and persuasive. There must be at least some common understanding. If the basic values are too discrepant, the evaluator has the option of not doing the study. Of course, those basic understandings are subject to prevailing conceptions of decency and justice in the society as a whole, and the evaluator has the option of drawing upon these larger social understandings.

That is not to say that the evaluator should be in total agreement with his audiences. Presumably, there are areas of disagreement or there would be no need for argument. Presumably, the audiences wish to learn something new or there would be no need for evaluation. But the evaluation proceeds from areas of agreement to those areas where agreement is problematic.

Quantitative Argument

The most popular approach to evaluation is the quantitative. Some see it as the very essence of rationality and scientific method. Many good evaluation studies have resulted from it–and many bad ones. Since this approach is taught in the graduate schools and promoted in the literature, there is little need to further extoll its virtues–they are many. In this section I would like to show that even quantitative methodology is essentially argumentation and is subject to similar considerations. Properly used, it can be a valuable tool of analysis; improperly used, it is dangerous.

Quantitative methodology is a body of mathematical methods and measurement techniques available to the evaluator. The utility of the methodology depends on similarities between the theoretical problems dealt with by the methodology and the substantive problems dealt with by the evaluator in the local setting. For his part, Cronbach (1974) has already determined that the fit on the theoretical and substantive problems is not a good one. The educational context is too complex.

A Rand Corporation mathematician (Strauch, 1976) examined the difficulties of quantitative methodology as it applies to policy studies, i.e., questions arising from the government decision-making process. According to Strauch, insofar as the methodology is mathematical, it is a self-contained system the structure of which is determined by the premises defining the system. Mathematical analysis is the exploration of that structure as it follows logically from the premises. The results are connected to the premises by logical inference. In the sense that their validity can be determined on the basis of that chain of reasoning, the results are "objective"—there is no need to appeal to the competence or judgment of the person who produced them nor to the audience to whom they are directed. The results are necessarily logical. In argumentation, by contrast, the results cannot be totally separated from the person who arrives at them.

The application of quantitative methodology to a substantive problem uses a mathematics model as a simplified representation of the problem. The results depend in part on the mathematical analysis—but equally on the fit between the model and the substantive problem. In the simplest applications, such as in physical science, the substantive problems are rigorously quantifiable. Experimental control enhances the ability of the evaluator to make the substantive problem conform to the mathematical model, i.e., randomness in statistical models. In such cases, the conclusions are "objective" in the sense that they are subject to independent verification on the basis of the logic and fit, without reference to the judgment of the person who produced them. However, the more behavioral or political the substantive problem the more difficult it is to define it unambiguously in mathematical terms. The links between the substance and the model become tenuous.

Strauch identifies the following components of such a quantitative study: *Formulation* involves defining the formal problem from the substantive problem, then finding a mathematical model for the formal problem. This is a process of reduction. *Analysis* involves computation within the mathematical context defined by the model. It results in mathematical statements. *Interpretation* means converting the statements back into the formal problem and finally interpreting these conclusions depends on *both* the logical validity of the analysis and the validity of the linkages. While the logical validity can be determined without reference to the subjective judgment of the analyst, the linkages cannot. They are founded upon the subjective judgments of the analyst. Both formulation and interpretation are subjective processes. Formulation requires reducing the substantive problem to something smaller that can be handled by the analysis and possibly adding some assumptions which make the analysis easier but may be questionable on substantive grounds, e.g., the independence of events.

Interpretation involves restoring the contextual considerations that have been eliminated and possibly adjusting for the simplifying assumptions. Both formulation and interpretation require considerable doses of intuitive judgment. Hence the conclusions are not really "objective" as claimed. (See the discussion of objectivity in a later section.)

The usual way of dealing with the subjective part of the methodology is to ignore it. For one thing it is not such a great problem in the natural sciences where quantitative methods have been so successful. Evidence of "objectivity" there is taken as proof of objectivity in other areas. When these links are challenged, it becomes clear enough that quite arguable premises underlie them.

Good insights are often derived from quantitative studies, but they usually result from the analyst making the right intuitive judgments rather than the right calculations. Those successes are often attributed to the quantitative methodology itself rather than to judgment. Critiques usually focus on the technical quality of the mathematical analysis rather than on the quality of judgments associated with formulation and interpretation. When quality of judgment is challenged, justification must rely on the kind of reasoning common to all argumentation.

One result of underplaying the role of judgment is what might be called "method-oriented analysis," according to Strauch. The analyst ignores the complexities of the context and plunges ahead with his favorite method. With superficial thought the methodology is applied in a straightforward manner as if there were no problems of fit. A few caveats are thrown in at the end suggesting that it is the readers' problem to decide whether the fit is a good one.

In its extreme form there is a school of thought which Strauch calls "quantificationism" which holds that quantification is a positive value in itself. A quantitative answer is always better than a qualitative one. Any problem can be reduced to a quantitative solution, and no problem can be properly understood until it is. Therefore quantitative methods should be applied to all problems. This position may be a straw man in that few people would really subscribe to it.

Such an attitude, which favors "scientific" methodology, is based on a reductionism that treats a phenomenon as an isolated system, develops a quantitative model for that system, and uses that model as a surrogate for the phenomenon. As suggested previously, reductionism may be one element of physical science not transferable to social phenomena.

The image the quantificationist projects is of a purveyor of objective "fact" based on hard data. He takes no personal responsibility for conclusions reached by his methodology since they are not of his making. He has simply uncovered them. He is merely reporting the results of his objective methods. He disdains qualitative data as subjective.

This attitude is close to what Polanyi (1958) described as "objectivism" in science. This is an attempt to define an objective method such that it relieves the observer of any responsibility for his findings. Polanyi contended, on the contrary, that the holding of a belief requires personal commitment and responsibility even in science. Objectivism has sought to represent scientific knowledge as totally impersonal.

Often quantificationism and objectivism also suit the decision-maker in that he may justify his decision by reference to a "scientific" finding. It may help him avoid personal responsibility. Attempts to quantify problems that are not quantifiable and to ignore the judgmental factors eventually distort decision-making.

Strauch suggests that one way to eliminate such distortion is to use quantitative methods as a *perspective* rather than a *surrogate* for the substantive problem. Accepting the mathematical model as a valid representation of the substantive problem means using it as a surrogate. Using the model by incorporating findings into *knowledge one already has* means using it as a perspective.

For most substantive problems, the audiences of the evaluation already have well-developed images of their own. The quantitative analysis may give the audiences an additional but not necessarily better or more valid insight into the problem. The interaction between one's own images and additional insights must take place in the heads of the audiences, the decision-makers or whoever. Using quantitative methodology as only one perspective reduces the problem of the fit between the model and the problem.

On the other hand, both the evaluator and the audiences must take more personal responsibility for the findings since they do not necessarily follow from the analysis. The conclusions cannot be justified entirely on the basis that they follow logically from the assumptions. Evaluation of individual assumptions must be supplemented by holistic evaluation of the total.

Quantitative argument, then, should always be used in conjunction with human judgment, and human judgment should be given the superior position. The implications for quantitative argument in evaluation are strong. Quantitative methodology should be seen to be based on human judgments and on intuitive reasoning and should be justified accordingly.

Qualitative Argument

In his paper on qualitative knowing, Campbell (1974) indicated that scientific knowing is dependent on common sense and that particular facts from either science or common sense are known only within the body of a great many other facts. "The ratio of the doubted to the trusted is always a very small fraction." Indeed, the knowledge of any detail is context-dependent, and, according to Campbell, qualitative knowing of "wholes and patterns" provides the context necessary for interpreting quantitative data. For example, generating alternative hypotheses requires familiarity with the local setting, a qualitative act.

Campbell believes that qualitative knowing has been neglected in favor of quantitative methods. At the same time he would prefer to see qualitative and quantitative methods used together to cross-validate one another. Quantitative methods, he believes, can provide insights that the qualitative do not, in spite of the prior grounding of the latter. Also, since all knowing is essentially comparative, he thinks qualitative techniques like case studies could be improved by experimental design considerations, which he would not see as being a part of quantitative methodology.

In rethinking the necessity and even the priority of qualitative knowing, Campbell (1975) has reconsidered the "anecdotal, single-case, naturalistic observation." Quantitative generalization will contradict such knowledge at some points but only by trusting a much larger body of such observations. In the classic paper on experimental design, Campbell and Stanley (1966), the case study was described as having no basis of comparison and hence providing no justification for drawing causal inferences.

Now Campbell has modified his position considerably, coming to believe that the case worker makes many predictions on the basis of his theory which he can disconfirm. The process is one of "pattern-matching" in which aspects of the pattern are matched against observations of the local setting. Campbell sees the single-shot case study as being a more secure basis of knowledge than he did in the past.

How is it in Campbell's view that we can know anything? He traces the current epistemological difficulties back to a quest for certainty in knowing. The effort to "remove equivocality by founding knowledge on particulate sense data and the spirit of logical atomism point to the same search for certainty in particulars" (Campbell, 1966). Certainty was to be established by defining "incorrigible particulars." This would result in unequivocally specifiable terms and in a "certainty of communication."

Campbell now sees this brand of positivism as not being tenable in either philosophy or psychology. Things out of context are not interpretable. But how can one still "know" something from a group of events which are each in themselves indeterminate? Campbell's answer is that this is achieved through "pattern-matching."

In events of cognition like binocular vision, the eyes recognize common objects by a process of triangulation. The more elaborate the pattern the more statistically unlikely a mistaken recognition becomes. Through memory various patterns can be compared. Pattern-matching itself Campbell sees as a trial and error process. This is essentially analogical thinking and Campbell sees it as being ubiquitous in the knowing process.

In fact, scientific theory is the most distal form of knowing, and the relationship between formal theory and data is one of pattern-matching with the error ascribed to the measurement of the data ("true" scores and "estimated" scores) except when it is agreed that the theory is in need of overhaul. There are two patterns to be matched, that of the theory, and that of the data. Acceptance or rejection of the theory is subject to some criterion of fit between the two. Actually, a theory is never rejected on the basis of its inadequacy of fit except when there is an alternative theory to replace it. It is the absence of plausible rival hypotheses that makes a theory "correct."

Campbell sees these considerations as directly relevant to program evaluation issues. "I believe that the problems of equivocality of evidence for program effectiveness are so akin to the general problems of scientific inference that our extrapolations into recommendations about program evaluation procedures can be, with proper mutual criticism, well-grounded."

If I understand his position correctly, Campbell is arguing that evaluation is a part of scientific inquiry and subject to similar epistemological concerns. However that may be, in this chapter at least, I have reversed the ground-figure relationship somewhat by treating science as an argument aimed at a universal audience and hence concerned with establishing long-term generalizations, and evaluation as an argument aimed at particular audiences dealing with context-bound issues.

In evaluation one may think of pattern-matching occurring not only in the evaluator's mind as he constructs his study and inspects the fit between his description of the program and the actual program itself, but also in the minds of the audiences as they compare the evaluation study to their own experience. The audiences themselves have images, memories, and theories of the program under evaluation. In using the evaluation as a perspective (in this case a verbal model), the audience matches its conception of the program to the evaluation. Where it attributes the error depends on the persuasiveness of the evaluation. The audiences thus serve as independent points of validation for the evaluation and must assume an active role in interpreting the evaluation and personal responsibility for the interpretation.

In Campbell's terms the basic pattern-matching process is analogical rather than logical (although the process must surely involve many forms of reasoning). In fact, one can go further than this. In an epistemology based on removing equivocality and establishing certainty of knowledge by defining "incorrigible particulars," deductive and inductive reasoning are the proper way of relating these particulars. Formal logic depends on unambiguous terms operating in a closed system.

To the extent that the terms are ambiguous and the system open (or not reducible to isolated subsystems), formal logic can be applied only argumentatively. The reasoning must include other varieties of thought or one must accept the fact that one cannot do rational analysis. Rational analysis is possible in evaluation but only rarely will it assume syllogistic form.[4]

Objectivity, Validity, and Impartiality Reconsidered

What does it mean to say that an evaluation study is "objective" or "valid?" Few concepts have been so confused and have caused so much mischief. Many people are reluctant to accept or believe qualitative evaluations simply because they are based on only one person's observations. Observations by one person are considered in and of themselves to be subjective and hence illegitimate for public purposes.

The crux of the confusion lies in misconceiving "objectivity." Scriven (1972) has written brilliantly about this confusion, tracing the unfortunate history of how objectivity has been defined. The theme of most definitions of objectivity is that there is something outside the mind that is verifiable through public or intersubjective agreement and that one can express or prove such things without influence from personal feelings. An evaluation which can do so is objective. But can one person's view ever be "objective"? The difficulty lies in confusing objectivity with procedures for determining intersubjectivity.

Scriven (1972) contended that there are two different senses in which objectivity is used–the quantitative and the qualitative. In the quantitative sense of the term, one person's opinion about something is regarded as being subjective–the disposition of one individual. Objectivity is achieved through the experiences of a number of subjects or observers. The common experiencing makes the observation public through intersubjective agreement. More formally, one might say that with a number of individuals one is more certain that one has properly represented the population–a sampling problem.

The qualitative sense of objectivity is quite different. It refers to the quality of the observation regardless of the number of people making it. Being objective means that the observation is factual, while being subjective means that the observation is biased in some way. Is it possible for one person's observations to be factual while a number of people's observations are not? Indeed it is. So an observation can be quantitatively subjective (one man's opinion) and also qualitatively objective (actually unbiased and true).

In fact, one might contend that the types of biases that affect the opinion of one person are somewhat different from those biases that plague group opinions. For example, an individual may succumb more easily to idiosyncratic viewpoints since he can hold only one perspective. On the other hand, there are social and cultural biases to which a group is more susceptible than is a particular person, e.g., jingoism. The individual's qualitative objectivity can be assessed by his previous track record on such matters and by his current self-interests. In any case, one who subscribes entirely to the quantitative notion of objectivity is not going to be satisfied with approaches like case studies.

How did the quantitative notion equating the number of people making an observation with its truth gain such ascendancy, even to the point of excluding qualitative objectivity? Scriven traces this distortion to psychology's attempt to root out introspectionism and philosophy's attempt to purge obscure metaphysics. Both tried to do so through the verification principle. Intersubjectivity became operationalized as *the* criterion for objectivity. In its extreme form the equating of objectivity with the quantitative notion of intersubjectivity was manifested in methodological behaviorism and in operationalism. But the fallacy of intersubjectivism pervades all fields.

Scriven cites the example of an evaluation of a television antenna in an electronics magazine in which the evaluator can see and report a better picture resulting from one of the tested antennas. Yet the evaluator apologizes for being "subjective" in his approach *since he did not use an instrument to measure decibel gain.* In fact, as Scriven notes, it is possible to get intersubjective agreement without instruments on the performance of electronic equipment, and it is the case that these pooled judgments of quality do not correlate highly with any instrument readings. Why then is an instrument reading objective while one person's judgment is subjective in the perception of this confused evaluator?

The reason is that the evaluator is only one person making the observation; and even though he knows he could have his observation confirmed by calling in his colleagues, he believes an instrument would be better because he can get even higher agreement among observers on the meter reading itself–even though the meter reading is not highly indicative of quality. In this case the quantitative notion of intersubjectivity has supplanted the quality of the perception.

In operational terms "measuring on a quantitative scale by mechanical means" becomes the indicator of truth because the interjudge reliability is higher, according to Scriven. Simultaneously, one has actually sacrificed validity for reliability because the meter reading, while reliable, is not a good indicator of picture quality. This is one of the common errors of evaluation–the substitution of instruments for direct observation of quality, the substitution of reliability for validity. And it is an error of the first magnitude.

From this idea–that what cannot be directly experienced by others cannot be taken seriously as science (intersubjectivism)–has developed the concept of objectivity as the externalization of all references so that multiple witnessing can be achieved, a gross oversimplification according to Scriven. In educational inquiry this has been manifested in equating objectivity with the ability to specify and explicate most completely all data collection procedures. Complete externalization and objectification permit replication, the hallmark of reliability. In education being objective has come to mean having a "valid" instrument–just as with the electronics evaluator.

What exists, in fact, are highly reliable instruments the validity of which is questionable. They do not always correlate highly with judgments of quality. The distortion of the intersubjectivist verification principle has resulted in equating objectivity with externalized, replicable procedures–even though these procedures may be infected by biases and hence be qualitavely subjective.

The identification of objectivity with a completely specifiable external procedure has another important effect. It relieves the evaluator of responsibility for the results and consequences of the evaluation. After all, if these "objective" instruments and procedures give these results, how can the evaluator be held liable? Science is to blame. Polanyi (1958) calls this position "objectivism." Objectivity in this sense comes to mean that observations are subject to independent verification without reference to the person who produced them.

Now it is not possible to specify all knowledge explicitly nor to verify it completely by independent-external procedures. Scriven contends that even in mathematical proofs in which the steps of the proof are reduced to the self-evident, intuition plays an inevitable and important role. Not only is intersubjective verification not a guarantee of truth, it is not necessary. Truth is an ideal which can only be *approximated* through an interplay of introspection and public verification.

Because of their complexity, many intuitive judgments can never be fully explicated. Yet conclusions may be no less true because of one's inability to explicate them. Agreement among many may be necessary for explaining the truth to someone else but it is not necessary for the truth itself.

How is it possible to establish the validity of a claim if one cannot separate it entirely from the person making the claim? One way is to check the reliability of the observer in previous instances and to check the observer's freedom from bias. These are not guaranteed to produce truth, but there are no guarantees. There are knowledge claims that are hybrids of the internal/external split, e.g., tendency statements, analogies, approximations, that are true yet are not the types of claims one usually associates with scientific statements, according to Scriven. He calls them "weak knowledge" claims and suggests they represent the type of knowledge available in the social sciences.

Such knowledge claims are manifested more as explanations than as predictions. Explanation and understanding are functions of the way information is coded in the mind. Explanation implies a person who is understanding the explanation. It does not exist by itself. The understanding is ultimately reducible to something familiar in the mind of the audience doing the understanding–or else it is not an explanation.

Similarly, unless an evaluation provides an explanation for a particular audience, and enhances the understanding of that audience by the content and form of the arguments it presents, it is not an adequate evaluation for that audience, even though the facts on which it is based are verifiable by other procedures. One indicator of explanatory power is the degree to which the audience is persuaded. Hence an evaluation may be "true" in the conventional sense but not persuasive to a particular audience for whom it does not serve as an explanation. In the fullest sense, then, an evaluation is dependent both on the person who makes the evaluative statement and on the person who receives it.

Prediction is not necessary to demonstrate understanding. Inferring an event from a correlation coefficient plus a few antecedent conditions is not necessary as a test of validity or objectivity. Rather, the basic reasoning pattern is closer to one of pattern-matching, of finding reasonable interpretations and explanations and understandings *within a given context.* The test of an explanation is not accuracy in predicting an event but whether the audience can see new relations and answer "new but relevant" questions.

Finally, about the question of objectivity one must conclude one of two things: either objectivity cannot be exclusively identified with an externalized procedure totally separated from the minds that produced the observations and comprehended them; or else a great deal of truth is subjective in character. In the first case, objectivity means something more than it is commonly taken to mean; in the second case, it means something less.

What about validity? One definition of validity is that it is based on objective procedures. Validity carries with it notions of being properly related to intent, of being correctly derived, and of being sanctioned by authority. In the narrow sense of quantitative objectivity, validity is equated with prediction–with checking the data against a criterion. But that assumes a single intent and assumes intersubjectivism as the verification principle. This is too narrow a procedure. Ultimately, says Cronbach (1971), validity is dependent on how the data are to be used and "utility depends upon values, not upon the statistical connections of scores."

If one cannot arrive at a single score presumably indicating validity, how is validity determined? Perhaps the best answer to the question is to examine the sources of invalidity. An evaluation may be invalid in a number of ways. One way is for the "facts and truths" upon which the evaluation is based to be wrong. Facts and truths are accepted without question by everyone. Other data must be determined by recognized data collection procedures, which are, in turn, sanctioned by a particular discipline and subject to public scrutiny. Often validity refers to using the accepted data collection procedures themselves, as Cronbach's article on test validation suggests.

Another way in which validity is at issue is in relating conclusions and interpretations to the data. As Cronbach asserts, it is not the test or the data collection procedures themselves so much as the interpretations that are valid or invalid. This is the validity of an inference. Is the inference correctly derived from the data and premises?

There is also the question of whether the interpretation can be properly applied to situations other than the one from which it was derived, since all generalizations are context-dependent. These concerns have been dealt with in experimental design somewhat systematically as threats to internal and external validity.

In qualitative studies it is more difficult to provide evidence of validity—which is not a sign that it does not exist. Demonstrating validity in naturalistic studies usually consists of confirming one kind of data with another kind. In proposing case studies of science education, Stake and Easley (1978) saw personal biases and past experience as the main threat to the credibility of the case studies. They proposed extensive tape recording of interviews, extensive use of direct quotations where possible, and reporting disagreements among respondents where they existed. People familiar with the local situation could read the written case to judge the accuracy of portrayal. Field workers would be keyed to "hints of inconsistency" for fruther pursuit. Contexts for observations would be documented and elucidated. Securing the observations of several participants about a particular issue or event was a way of "triangulating" what actually happened.

Most of these threats to validity are seen from the perspective of a universal audience. But there is another way of looking at validity in evaluation—whether the evaluation is valid for particular audiences. After all, validity is always concerned with purpose and utility for someone. If the evaluation is not based on values to which the major audiences subscribe, these audiences may not see it as being "valid," i.e., relevant to them in the sense of being well-grounded, justifiable, or applicable. The evaluation may simply miss the main issues as far as particular audiences are concerned. At the same time the evaluation may be valid in the sense that the facts are correct and the inferences from the data correctly derived. From a particular audience's perspective, the premises may be the wrong ones.

An evaluation can also be invalid in this secondary sense if the argument forms employed are wrong. For example, in this society "means-ends" arguments, particularly cost-effectiveness arguments, are particularly potent. If one were to employ an argument based on maximizing excellence instead of choosing the best available alternative, it might carry little weight although being equally true and valid from the perspective of the universal audience. So validity can apply to evaluation in rather different ways. (The debate between Glass and Scriven in Appendix A is over the form of the argument as much as anything.)

It is also the case that the more "naturalistic" the evaluation, the more it relies upon its audiences to draw its own generalizations (external validity). For example, a case study may be interpreted in different ways by each reader, since each reader has her own universe of cases in her mind for comparison. The reader can see similarities and differences based on her own experience and can draw her own interpretations.

Conceiving the process of generalization in this way alters even the first sense in which validity is used. The evaluator is still responsible for ascertaining and reporting "true" facts and statements, but part of the interpretation is beyond him. Since, as Cronbach says, the ultimate issue is the validity of the interpretation, which only the reader knows for sure, the audiences must assume considerable responsibility for the validity of their own interpretations. The evaluator must ultimately assume rational processes in the thinking of the audiences.

As Ennis (1973) noted, internal validity and external validity refer to rather different phenomena. External validity is concerned with the generalizability of general causal statements. Internal validity bears on specific causal statements that do not entail generalizing to new cases. Generalizing always assumes that one knows the relevant laws involved in extrapolating into new realms. An internally validity study, by contrast, only claims causality in the past within the specific circumstances. It claims no extrapolation and is hence less dependent on outside assumptions.

However, neither specific causal statements nor general causal statements follow perfectly logically from observations, even in the best experimental designs. Some empirical assumptions are needed even in the tightest design. In addition, identifying a particular event as a cause inescapably involves a judgment of responsibility that a particular event is responsible for the effect, according to Ennis. This ascription of responsibility requires much background knowledge and a value judgment. It involves a probable assignment of praise or blame and suggests a place for intervention.

Most evaluators would assume responsibility for specific causal statements that "x caused y" in this study (internal validity), although this in itself necessarily involves a set of assumptions. But some would refer the generalizability of the findings to the audiences' judgments, since generalizability is based on outside information which the audiences but not the evaluator may have. The audiences might make some of the responsibility ascriptions based on their own background knowledge and values. Some evaluators, particularly naturalistic ones, might argue that this would ultimately result in superior generalizations.

There is yet a further related problem with objectivity. Is it really sufficient to say that an evaluator is objective? If objectivity is taken in the commonly used sense of employing an externalized, specifiable procedure which produces replicable results, then it is certainly an insufficient criterion for an evaluation. The administration of standardized achievement tests is a totally externalized, specifiable procedure which produces replicable results. At the same time such tests are thought to be highly biased in many ways, particularly toward minority groups. In this sense, one has an objective but biased instrument. In fact, one can produce an instrument in which the bias is in the other direction. (To further confound matters, if racial discrimination is the intent of such an instrument, one could have an objective, valid instrument for that purpose.)

An evaluation must be free from distortion and bias (qualitatively objective), and being externalized, specifiable, and replicable does not sufficiently address possible biases. Even qualitative objectivity is insufficient for evaluation, for it carries the aura of neutrality. People being evaluated do not want a neutral evaluator, one who is unconcerned about the issues. A person on trial would not choose a judge totally removed from his own social system.

Being disinterested does not give one the right to participate in a decision that determines someone's fate to a considerable degree. Knowledge of techniques for arriving at objective findings is inadequate. Rather, the evaluator must be seen as a member of or bound to the group being judged, just as a defendent is judged by his peers. The evaluator must be seen as caring, as interested, as responsive to the relevant arguments. He must be impartial rather than simply objective.

The impartiality of the evaluator must be seen as that of an actor in events, one who is responsive to the appropriate arguments but whom the contending forces are balanced rather than nonexistent. The evaluator must be seen as not having previously decided in favor of one position or the other.

The evaluator may resport to objective criteria to resolve the issues; but when his own impartiality is at stake, it is not enough that he give evidence of objectivity. He must give evidence of his impartiality by showing how he has acted contrary to his own interests in the past.

Evaluative Discourse: The Good Life (Along the San Andreas Fault)

It has been several weeks since I began this chapter. The great Los Angeles earthquake has not yet come. Beautiful day succeeds beautiful day, each one much like the last; so it seems tomorrow must be like today, a pleasant dream extending indefinitely.

Each day that passes makes the quake seem less likely than before. Yet if it is to occur this year, it should become *more* likely. I reason that the time I have remaining here is only a small fraction of the coming year, so the chances of the quake coming now are less than for the entire year of the prediction. I reason that even if the quake should come, the effects will not be disastrous. In addition, the Midwest is racked by tornadoes. Besides, would many of the smartest men in the country, including the seismologists, live here if the danger were so great? I feel reassured. My anxiety lessens.

Meanwhile within the last few days, the *New York Times Magazine* heightens the drama in its Bicentennial edition (July 4, 1976). As symbolic of "America at 200," it features a report on "The Good Life (along the San Andreas Fault)." On the cover is a painting of a fragment of a freeway jutting out into the empty ocean, the remains of Los Angeles after the next earthquake. The article begins with a six-paragraph scenario of the effects of the anticipated quake.

Those who live on top of the nine-mile deep fault have their own reasons for living there. As his backyard crumbles away daily, a postal worker, who has three cars, would like to move but cannot sell his house. A ranch manager, who finds life better in California than anyplace he has ever lived, explains, "I'm not leaving. Is there any place that doesn't have some catastrophe?" For some, precariousness itself makes being here all the more precious. A dropped-out investment counselor living on the fault says, "You're living on a crisis point. Everything you have can be taken away from you at any time."

These are not the reasons I would give but they may be right. Each person is free to weigh his own reasons. Each is free to make his own choices. So it must be when faced with such uncertainty of knowing. Judgments cannot be based on an irrefutable reality. Even when earthquakes are much more predictable, there will remain room for choice in how to respond. In social decision-making certainty seems remote if not impossible.

Faced with such difficulty in arriving at an irrefutable reality, there are those who try to force simplicity atop the complexities of life. They insist on pretending there is agreement where there is none, whether of facts or of values. Often in positions of power, they impose simplified definitions of reality for the sake of action. Yet, no matter how widely accepted the simplification, reality is still there. Whatever twenty-one million Californians believe, the great earthquake will come eventually.

The alternative is not necessarily a descent into irrationality. If opinions cannot be indisuptably based, neither must they be regarded as entirely arbitrary, as being merely "value judgments." Such a classification limits knowledge to that which is clear, distinct, and unambiguous. This distinction establishes a schism between objectively true theoretical knowledge on the one hand and action based on irrational motives on the other. It culminates in designating as irrational those who do not agree with one's perspective. Classifying people as irrational justifies ignoring their opinions and perhaps their dignity and interests. It even legitimates using suggestion and force on them.

The alternative is to treat all men as rational. Between the conservative authoritarianism of tradition and the liberal authoritarianism of scientism, between the certainty of fanaticism and the evasion of responsibility of skepticism lies rational deliberation. One must take seriously the opinions of other people and engage them in serious discourse. This is the realm of argumentation and the proper sphere of evaluation.

The starting point is that groups of people adhere to opinions with variable intensity, and that these opinions can be put to the test of serious discourse. Even facts and values may be so considered. Rational discourse consists of giving reasons, although not compelling reasons. In the realm of action, where few things are clear and distinct, motivation can be rational. Practice can be reasonable.

The evaluator must engage his audiences in a dialogue in which they are free to employ their reasoning. This means that the audiences must assume personal responsibility for their interpretation of the evaluation since the reasoning presented to them is neither completely convincing nor entirely arbitrary. This means that the evaluator must also assume personal responsibility for his judgments since he cannot hide behind blind method. Both must exercise their natural reason.

NOTES

[1] For this distinction and many other ideas in this paper, I am indebted to Perelman and Olbrechts-Tyteca's excellent modern work on argumentation *The New Rhetoric: A Treatise on Argument*, Scranton, PA: Univ. of Notre Dame Press, 1969.

[2] At the end of his masterpiece on inductive logic, Mill considers the logic of a "practice" or "art." "There must be some standard by which to determine the goodness or badness, absolute and comparative, of ends, or objects of desire. And whatever that standard is, there can be but one; for if there were several ultimate principles of conduct, the same conduct might be approved by one of those principles and condemned by another; and there would be needed some general principle, as umpire between them" (John Stuart Mill, *A System of Logic*, New York: Harper, 1893 [8th Edition]).

This leads Mill to impose a single universal standard by which to judge practical affairs, for the only alternative is by "supposing a moral sense or instinct" or "intuitive moral principles." General ethical principles can be known only by induction. Since inductive certainty presupposes a uniformity of nature, the resultant psychology is deterministic. Morality is natural since only a naturalistic assessment will allow scientific methods of proof. Hedonistic utilitarianism is the basis.

In a sense, Mill was preventing disagreement over moral issues since it is always possible to reach opposite conclusions when there is no previous agreement on a criterion. The result of this reasoning is utilitarian calculation which conflates all human desires into a single configuration and satisfies them by the criterion of maximum total satisfactions derived. The judging is done by an "impartial spectator," who in modern times demonstrates his impartiality by employing "objective" techniques of analysis.

[3] Kelly (1980) has pointed out the hidden premises in my own arguments here. It is impled that the evaluator acts to persuade the audience of a point of

view because it is "true," and he has some way of establishing this. Truth can be held with varying intensity, however. Kelly also claims, correctly I think, that I am making the Aristotelian distinction between theoretical argument that leads to truth and practical argument that leads to reasoned action.

I am less certain about his claim that evaluation persuades someone to act rather than persuades them that something is the case. Action is the ultimate goal of evaluation, but there are so many other considerations involved in action that it seems unlikely the evaluator would be able to assess, or even identify the major contingencies. It seems to me that evaluation persuades as to the worth of something. Under some circumstances this may be a course of action, but ordinarily the action entails additional considerations.

4. For an extended analysis of an evaluation as argumentation, see Appendix A. For an analysis of "naturalistic evaluation," see Appendix B.

5

COHERENCE AND CREDIBILITY
The Aesthetics

The Drunken Driver

> Humankind lingers unregenerately in Plato's cave, still reveling, its age-old habit, in mere images of the truth.
>
> Susan Sontag, *On Photography*, 1977: 1

Consider two different images of the drinking driver. One may imagine the ordinary social drinker who happens to overindulge, and who, missing a stop sign, is detained by the police, thereby getting into trouble. Or imagine the drunken driver, one who is habitually drunk, a reeling, stumbling, insensate hazard to everyone on the road, including himself. The image that one constructs of the driver who drinks has much to do with the recommendations for action that one might embrace as a means of curtailing drinking drivers.

AUTHOR'S NOTE: This chapter was published as "Coherence and Credibility: The Aesthetics of Evaluation," *Educational Evaluation and Policy Analysis*, Vol. 1, No. 5, September-October, 1979, pp. 5-17. Copyright 1979, American Educational Research Association, Washington, DC. Reprinted by permission. Helpful comments were provided by Gene Glass, Jo Friedman, Donald Hogben, Rochelle Mayer, John Nisbet, and Helen Simons.